-6

## Problem 1: True or False

Please indicate whether the following statements are true (T) or false (F).

1. (2 pts) For two linear regression models A and B, if A is simpler than B, then A will have a better performance than B on the testing set.
   Answer: F

2. (2 pts) Usually we do not use regression models for classification. But there exists special cases. For example, Logistic regression is a model used for regression, and it can be used for 2-class classification problem.
   Answer: T

3. (2 pts) Suppose we have a dataset. It contains 900 images of class "cat" and 100 images of class "dog". If we train a classifier which achieves 85% accuracy on this dataset, then it is a good classifier.
   Answer: F

4. (2 pts) The perceptron algorithm that we learned in class makes use of a variable learning rate, which decreases as the algorithm progresses.
   Answer: F
   maximization ↓ learning rate fixed.

5. (2 pts) In the primal version of SVM, we are minimizing the Lagrangian with respect to $w$. In the dual version, we are ~~minimizing~~ the Lagrangian with respect to $\alpha$.
   Answer: F
   ~~minimizing~~ ↑

6. (2 pts) The $k$-means algorithm is used for unsupervised learning. As it is unsupervised, we do not have to specify the number of clusters $k$ before we start running the algorithm. The value $k$ will be learned automatically from data in an unsupervised manner.
   Answer: ~~O~~ T
   Elbow method :→ Pg 1. we could choose k many training examples randomly

7. (2 pts) Gradient descent may get stuck in local minimum points, but EM does not.
   Answer: ~~F~~ F. see comparison with K means, EM slides.

8. (2 pts) We can use a held-out validation set to tune our model's hyper-parameters, such as the regularization coefficient $\lambda$ in logistic regression or SVMs.
   Answer: T

9. (2 pts) Convergence of the BackPropagation algorithm is generally not guaranteed, unless the error surface is convex.
   Answer: T

10. (2 pts) Assuming that you are not concerned with the training time, when using a deep learning network it is best to include as many hidden units as possible, so the training error can be reduced as much as possible.
    Answer: ~~F~~ T
    overfitting.

justification: 2

Training error is reduced, but.
Validation error is greater.

## Problem 2: Multiple Choice Questions

1. (3 pts) After SVM learning, each Lagrange multiplier $\alpha_(x, y)$ takes either zero or non-zero value. What does it indicate in each situation?

   - ( ) A non-zero $\alpha_(x, y)$ indicates the data point $(x, y)$ is a support vector, meaning it touches the margin boundary.
   - ( ) A non-zero $\alpha_(x, y)$ indicates that the learning has not yet converged to a global minimum.
   - ( ) A zero $\alpha_(x, y)$ indicates that the data point $(x, y)$ has become a support vector data point, on the margin.
   - ( ) A zero $\alpha_(x, y)$ indicates that the learning process has identified the class for data point $(x, y)$.

2. (3 pts) Which of these is NOT a good learning rate decay scheme? Here, $t$ is the epoch number, and $\alpha_0$ is the initial learning rate.

   - ( ) $\alpha = \frac{1}{1+2\times t}\alpha_0$.
   - ( ) $\alpha = e^t\alpha_0$.
   - ( ) $\alpha = 0.95^t\alpha_0$.
   - ( ) $\alpha = \frac{1}{\sqrt{t}}\alpha_0$.

3. (3 pts) Which statement of the following is INCORRECT *w.r.t* a recommendation system?

   - ( ) We subtract the average ratings of each user to eliminate user bias?
   - ( ) The user rating matrix is usually sparse.
   - ( ) Training the K-Nearest Neighbors method is very time consuming.
   - ( ) The idea of collaborative filtering is to discover similar users to the user-of-interest, so that his / her rating can be predicted based on the similar users' ratings.

4. (3pts) Thinking about unsupervised learning and the k-Means and expectation maximization (EM) algorithms, which one of the following statements is true

   - ( ) K-Means assigns a probability to the membership of each example to each cluster.
   - ( ) The K-Mean salgorithm obtains a global optimal solution for the partition of a dataset by minimizing the square distance between examples and their nearest centroid.
   - ( ) K-Means using the euclidean distance is a particular case of the EM-algorithm when we are fitting $K$ gaussian distributions with the same variance.
   - ( ) EM algorithm has the same computational cost regardless of the number of parameters that have to be estimated for the probability distribution that we are fitting to the random variables.

5. (3 pts) Which of the following statements about generative adversarial networks and recurrent neural networks is INCORRECT?

   - ( ) The generative adversarial network is a particular case of convolutional neural network.

see blank page for $\phi_2(y)$ work

**Problem 3: Kernel Method**

(3 pts) Let $k_1$ and $k_2$ be (valid) kernels; that is, $k_1(\mathbf{x},\mathbf{y}) = \phi_1(\mathbf{x})^T\phi_1(\mathbf{y})$ and $k_2(\mathbf{x},\mathbf{y}) = \phi_2(\mathbf{x})^T\phi_2(\mathbf{x})$. Show that $k = k_1 + k_2$ is a valid kernel by explicitly constructing a corresponding feature mapping $\phi(\mathbf{x})$. Hint: $\phi(\mathbf{x})$ is represented by both $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$.

A function is a kernel function if

① $K(x,y) = K(y,x)$

② $K_{ij} = K(x^{(i)}, x^{(j)})$.

is positive semidefinite.

$k_1(x,y) = \phi_1(x)^T\phi_1(y)$

$k_2(x,y) = \phi_2(x)^T\phi(x)$

from slide 6,

$K(x,x') = \phi(x)\cdot\phi(x')$

$K = \phi_1(x)^T\phi_1(y) + \phi_2(x)^T\phi_2(y)$

$= \phi(x)^T\phi(y)$

① & ② are met

with $\phi(x)$ 5 being with $\phi(y)$

mapping to $\phi_1(y)$ + $\phi_2(y)$

$\phi(x)^T = \phi_1(x)^T + \phi_1(x^T)$

-2.5

4

## Problem 4: Support Vector Machine

You have trained a simple linear SVM from a large collection of data. Now you would like to explore the trained model a little further.

1. (4 pts) You found the margin boundaries are $3x_1 + 12x_2 + 4x_3 + 1 = 0$ and $3x_1 + 12x_2 + 4x_3 + 3 = 0$. What is the decision boundary? What is the size of the margin? What are the values of $\theta$ and $\theta_0$ of the decision boundary, respectively? Hint 1: $3^2 + 4^2 + 12^2 = 13^2$. Hint 2: the decision boundary is between the margin boundaries. Hint 3: the distance between two parallel lines $Ax + By + C_1 = 0$ and $Ax + By + C_2 = 0$ is $\frac{|C1-C2|}{\sqrt{A^2+B^2}}$.

$3x_1 + 12x_2 + 4x_3 + 1 = 0$

$3x_1 + 12x_2 + 4x_3 + 3 = 0$.

So "

size of the margin.

$z = \begin{bmatrix} 3 \\ 12 \\ 4 \end{bmatrix}$ ✓

$= \frac{3 - 1}{\sqrt{13^2}}$

$= \frac{2}{13}$ ✓

$|z| = 13$.

decision boundary

$= 13z + 2$

$= 3x_1 + 12x_2 + 4x_3 + 2$ ✓

where $\theta = 13$, $\theta_0 = 2$

−1

2. (4 pts) Next, in the same training dataset, you found the following points of the form $(x_1^{(t)}, x_2^{(t)}, x_3^{(t)}, y^{(t)})$: $((-1, 1, -2), +1)$, $((0, 1, -4), +1)$, $((-1, 1, -3), -1)$, $((0, 0, 0), -1)$. Is this dataset linearly separable? Clearly explain why. Which of these points are support vectors?

Proximity +- one another

$$\begin{bmatrix} 3 \\ 12 \\ 4 \end{bmatrix} [-1, 1, -2) = -3 + 12 - 8 = 1$$

$$\begin{bmatrix} 3 \\ 12 \\ 4 \end{bmatrix} [0, 1, -4) = -4$$

$$\begin{bmatrix} 3 \\ 12 \\ 4 \end{bmatrix} [-1, 1, -3) = -3 + 12 - 12 = -3.$$

This is not linearly separable as the z value do not categorize correctly

$y = +1$ $\boxed{1, -4}$ $\quad y = -1$ $\boxed{-3, 0}$

The support vectors are those farthest away from the decision boundary, that is, where the calculated values are not close to 0: In this case, we see that

$-2, 5.$ → see blanked page after problem 5

1 −2+2−1

2 +2+2+2

## Problem 5: Convolutional Neural Network

Suppose we have an image of size $4 \times 4$, written as

$$\begin{bmatrix} 1 & 2 & -2 & -1 \\ -1 & -1 & 2 & 0 \\ 1 & 1 & -1 & 1 \\ 0 & 0 & 2 & 1 \end{bmatrix}$$

2+2+2+2

We have a convolutional filter of size $2 \times 2$, written as

$$\begin{bmatrix} 1 & -1 \\ -2 & 1 \end{bmatrix}$$

−2+1−2

0

(2 pts) Calculate the output feature map between the input image and the convolutional filter. Assume the stride is 1. Can you explain the resulting feature map (which image regions activate the filter)?

−1 −2 −2 −1

2+0+2+1

1 −1 0 ..

1 +1+2

−1 −4

+1

$$\begin{bmatrix} 0 & 8 & -5 \\ -1 & -6 & 5 \\ 0 & 4 & -5 \end{bmatrix}$$

assuming that ordered

$$\begin{bmatrix} x_{11} & x_{12} & \\ x_{21} & & \\ & & x_{44} \end{bmatrix}$$

can tell that regions

$$\begin{bmatrix} x_{12} & x_{13} \\ x_{22} & x_{23} \end{bmatrix}$$

activate filter positively

2 $\begin{bmatrix} x_{32} & x_{33} \\ x_{42} & x_{43} \end{bmatrix}$

(2 pts) Given the computed feature map, now we perform max pooling on the resulting feature map. Suppose the max pooling window size is $2 \times 2$, and the stride for the pooling window is 1. Calculate the result of max pooling.

pooling window = $\begin{bmatrix} 1 & -2 \\ -3 & -2 \end{bmatrix}$

$1 - 7 = 1$

$13 - 11 = 2$

$-7 + 4 = -3$

$-11 + 9 = -2$.

$-2$